

Marie CHAVENT (MdC)

Logiciels :

1) ClustOfVar

Descriptif : ClustOfVar est un package R dédié au clustering de variables. Les variables peuvent être numériques, catégorielles ou mixtes (un mélange des deux). Le package propose un algorithme de clustering hiérarchique et un algorithme de type k-means. Le centre d'une classe de variables est ici une nouvelle variable synthétique, et non un centre de gravité comme pour le clustering de Ward ou des k-means classique (en clustering des observations). Cette variable synthétique d'une classe est la première composante principale de la méthode d'ACP de données mixtes appelée PCAmix. L'homogénéité d'une classe est alors définie comme la somme des carrés des corrélations (pour les variables numériques) et des rapports de corrélations (pour les variables catégorielles) à la variable synthétique de la classe (qui est toujours quand à elle numérique). Le clustering de variable peut-être utilisé comme outil d'exploration de la structure des corrélations entre les variables ou encore comme outil de réduction de dimension permettant de diminuer la redondance entre les variables. C'est en ce sens un outil alternatif à l'ACP pour réduire la dimension avant d'appliquer d'autres traitements statistiques.

Brevet :

Nombre de lignes : 1700

OpenSource : oui

Langage : R

Bibliothèques :

Postraitement :

Parallélisation :

Page web :

2) PCAmixdata

Descriptif : PCAmixdata est un package R dédié à l'analyse multidimensionnelle de données mixtes (mélange de données numériques et catégorielles). Trois méthodes sont principalement implémentées dans le package. La méthode PCAmix inclut l'ACP (analyse en composantes principales) pour des données numériques et l'ACM (analyse des correspondances multiples) pour des données catégorielles. La méthode PCArot permet d'effectuer une rotation de type varimax des composantes principales de PCAmix. Et enfin, MFAmix permet d'effectuer une AFM (analyse factorielle mixte) pour des données de type multi-tableau, avec de la mixité à l'intérieur des groupes de variable. Des fonctions de prédictions des composantes principales extraites par les trois méthodes sont également implémentées, permettant de calculer les scores de nouvelles observations en validation croisée par exemple.

Brevet :

Nombre de lignes : 2500

OpenSource : oui

Langage : R
Bibliothèques :
Postraitement :
Parallélisation :
Page web :

3) divclust

Descriptif : DIVCLUS-T est une méthode descendante de clustering hiérarchique basée sur une approche de bi-partitionnement monothétique permettant de lire le dendrogramme comme un arbre de décision. Comme pour la méthode ascendante de Ward, le critère de qualité d'une partition est l'inertie intra-classe. A chaque étape de l'algorithme, une classe est divisée en deux selon une question binaire (définie sur une seule variable), mais on choisit dans l'ensemble de toutes les questions binaires possibles celle qui induit la partition de plus petite inertie intra-classe. Cette inertie intra-classe quant à elle est calculée de manière classique sur toutes les variables. Ainsi si n observations sont décrites par p variables dans un noeud par exemple, il y aura $p*(n-1)$ questions binaires possibles et donc $p*(n-1)$ bi-partitions des n observations à évaluer. La bi-partition ayant l'inertie intra-classe la plus petite est alors retenue et la question binaire caractérise cette division. La méthode DIVCLUS-T construit ainsi une hiérarchie indicée comme celle de Ward, mais le dendrogramme fournit pour chaque classe une règle de décision, condition nécessaire et suffisante d'appartenance à la classe. La méthode DIVCLUS-T permet en outre le clustering de données mixtes (numériques et catégorielles).

Brevet :
Nombre de lignes : 1200
OpenSource : oui
Langage : R
Bibliothèques :
Postraitement :
Parallélisation :
Page web :

Ressources de calcul :

Chercheurs - Ingénieurs - Doctorants – Post-doctorants – Collaboration extérieures :

Pour ClustOfVar et PCAmixdata:

- Marie Chavent et Jérôme Saracco (IMB et Equipe CQFD inria)
- Vanessa Kuents et Amaury Labenne (Irstea)
- Benoit Liquet (Isped et Université du Queensland)

Pour divclust : Marie Chavent et Marc Fuentes (Ingénieur SED Inria Bordeaux)

Plate-formes de développement :

Dépôt SVN ou GIT sur une plate-forme ? Autre ?

Pour ClustOfVar :

- sur github : <https://github.com/chavent/ClustOfVar>
- sur le cran : <http://cran.r-project.org/web/packages/ClustOfVar/index.html>

Pour PCAmixdata :

- sur github : <https://github.com/chavent/PCAmixdata>
- sur le cran : <http://cran.r-project.org/web/packages/PCAmixdata/index.html>

Pour divclust :

- sur la forge inria : https://gforge.inria.fr/frs/?group_id=4729

Besoins :

- 1) Formation GIT et Github**
- 2) Formation outils de débogage pour un code R**
- 3) Comment bien gérer les différentes versions d'un logiciel**
- 4) Comment améliorer les performances d'un code R**